

ORIGINAL ARTICLE

Development and validation of an international appraisal instrument for assessing the quality of clinical practice guidelines: the AGREE project

The AGREE Collaboration*

Qual Saf Health Care 2003;12:18–23

Background: International interest in clinical practice guidelines has never been greater but many published guidelines do not meet the basic quality requirements. There have been renewed calls for validated criteria to assess the quality of guidelines.

Objective: To develop and validate an international instrument for assessing the quality of the process and reporting of clinical practice guideline development.

Methods: The instrument was developed through a multi-staged process of item generation, selection and scaling, field testing, and refinement procedures. 100 guidelines selected from 11 participating countries were evaluated independently by 194 appraisers with the instrument. Following refinement the instrument was further field tested on three guidelines per country by a new set of 70 appraisers.

Results: The final version of the instrument contained 23 items grouped into six quality domains with a 4 point Likert scale to score each item (scope and purpose, stakeholder involvement, rigour of development, clarity and presentation, applicability, editorial independence). 95% of appraisers found the instrument useful for assessing guidelines. Reliability was acceptable for most domains (Cronbach's alpha 0.64–0.88). Guidelines produced as part of an established guideline programme had significantly higher scores on editorial independence and, after the publication of a national policy, had significantly higher quality scores on rigour of development ($p < 0.005$). Guidelines with technical documentation had higher scores on that domain ($p < 0.0001$).

Conclusions: This is the first time an appraisal instrument for clinical practice guidelines has been developed and tested internationally. The instrument is sensitive to differences in important aspects of guidelines and can be used consistently and easily by a wide range of professionals from different backgrounds. The adoption of common standards should improve the consistency and quality of the reporting of guideline development worldwide and provide a framework to encourage international comparison of clinical practice guidelines.

*See end of article for contributors

Correspondence to:
Dr F Cluzeau, Department
of Public Health Sciences,
St George's Hospital
Medical School, Cranmer
Terrace, London
SW17 0RE, UK;
f.cluzeau@sghms.ac.uk

Accepted for publication
26 June 2002

Clinical practice guidelines are now a common feature of clinical practice and are of interest worldwide. They are expected to facilitate more consistent, effective and efficient medical practice, and improve health outcomes.¹ Governments, professional associations, and healthcare organisations are increasingly sponsoring the development and dissemination of clinical guidelines.² There is also a growing number of guidelines developed by European or international groups.

Although the principles for the development of sound guidelines are well established,^{3–5} many published guidelines fall short of the basic quality criteria identified in two recent studies.^{6,7} Defining the quality of guidelines is not straightforward. In principle a "good" guideline is one that eventually leads to improved patient outcome. It needs to be scientifically valid, usable, and reliable. However, this evidence is rarely available. Often the best that can be expected is some information on whether the guideline producers have attempted to minimise all the biases that can occur in the complex process of creating a guideline and how well this is reported.

As the number of published guidelines proliferates, there have been calls for the establishment of internationally recognised standards to improve the development and reporting of clinical guidelines.⁸ Moreover, there is a pressing need for internationally recognised criteria that are valid, reliable, and useful for various assessment purposes in different countries, both for guideline developers and clearing houses as well as individual users of guidelines.

In response, an international group of researchers from 13 countries—the Appraisal of Guidelines, REsearch and Evaluation (AGREE) Collaboration—has developed and validated a generic instrument that can be used to appraise the quality of clinical guidelines. The AGREE instrument is designed to assess the process of guideline development and how well this process is reported. It does not assess the clinical content of the guideline nor the quality of evidence that underpins the recommendations. In this paper we report the development and validation of the AGREE instrument.

METHODS

A multi-staged approach was used that included an item generation, selection and scaling process, and field testing and refinement procedures.

Item generation, selection, and scaling

To develop the framework for the instrument, quality was defined as the confidence that the biases linked to the rigour of development, presentation, and applicability of a clinical practice guideline have been minimised and that each step of the development process is clearly reported. We considered the following five theoretical quality domains:

- scope and purpose;
- stakeholder involvement;
- rigour of development;
- clarity and presentation;

Box 1 Participating countries, and selection criteria for guidelines and appraisers

Participating countries: Canada, Denmark, England, Finland, France, Germany, Italy, The Netherlands, Scotland, Spain, Switzerland (England and Scotland were considered separately because they have independent guideline programmes).

Selection criteria for guidelines:

- guidelines published between 1992 and 1999
- preferred disease areas: asthma, breast cancer, and diabetes
- documents that contain specific recommendations for clinical practice (excluding systematic reviews or service documents)

Selection criteria for appraisers:

- broad range of professions including clinical experts, nurses, researchers and policy makers
- different healthcare settings including primary care, secondary care, teaching hospitals
- excluding members from guideline development group

- applicability.

A small working group (FC, JB, RG, PL) generated an initial list of 82 items from validated appraisal instruments and relevant literature⁶⁻¹² that addressed these domains. The working group examined the list for coverage, overlap and content validity, and reduced it to 34 items. The list and a user guide describing the items were pretested on two Dutch and two English guidelines and refinements were made in response to the comments received.

The refined list and user guide were then circulated to all the AGREE partners and to 15 international experts for their views on the clarity, comprehensiveness, relevance, and ease of use. In addition, the AGREE partners were asked to apply the instrument to two guidelines each. The feedback from this process led to reformulation of ambiguous items and removal of overlapping and value laden items. The result was the first draft instrument comprising 24 items grouped into the five domains identified in the development phase. We also modified the user guide to reflect changes made to the items. A 4 point Likert scale was used to score each item (1=strongly disagree, 2=disagree, 3=agree, 4=strongly agree). A 3 point scale (1=not recommend, 2=recommend with provisos or modifications, 3=strongly recommend) was used to score an overall judgement on whether the guideline ought to be recommended for use.

Field testing and refinement

The AGREE collaborators field tested the instrument following a research protocol that covered selection criteria for the guidelines, methods for recruiting appraisers, and time scales (box 1). Each country coordinated the appraisal of at least seven guidelines. Each guideline was assessed independently by four appraisers and, where possible, each appraiser assessed two guidelines. The appraisers received a standard letter with instructions on how to complete the instrument. Most used an English version of the draft AGREE instrument. If necessary, the materials or the user guide only were translated to ensure appraisers' understanding of the items. Feedback on the instrument, user guide, and the appraisal process was solicited with a standard letter, translated into a national language where necessary.

The field test was conducted in winter 1999–2000 with the 24-item draft instrument. For this phase, 100 guidelines from 11 countries (mode=8, range 7–22) were evaluated by 194 appraisers. The results of this field test were reviewed at an AGREE workshop in spring 2000 and the instrument and user guide were refined in response to the results. The final version

of the instrument underwent further field testing in autumn 2000. In this phase a random sample of three guidelines per country from the original 100 were assessed by 70 newly recruited appraisers.

Data analysis

Mean item scores for each guideline were calculated by averaging the scores across the four appraisers. Standardised domain scores for each guideline were calculated by summing scores across the four appraisers and standardising them as a percentage of the possible maximum score a guideline could achieve. Mean item and standardised domain scores were used in the analyses unless otherwise noted below.

To guide the refinement of the instrument from the draft version to the final version, a principal components analysis was undertaken with data from the first field test. The mean item scores for each of the 100 guidelines were included in the analysis, with the eigen value limit set at 1 and the criteria for the minimum loading score set at 0.52.^{13 14}

Final instrument properties

Reliability

Two measures of reliability were conducted:

- (1) Using mean item scores, the Cronbach α coefficient was calculated to measure internal consistency of each domain of the final instrument.¹⁵
- (2) Intraclass correlations (ICC) were calculated to assess the reliability within each domain. ICCs based on single appraisers' ratings and the means of two, three, and four appraisers were calculated.¹⁶

Validity

Several measures of validity were considered:

- (1) Face validity: appraisers' attitudes about the instrument and user guide were collected by questionnaire and used to assess face validity.
- (2) Construct validity: three hypotheses were considered for tests of construct validity:
 - (a) Established guideline programmes have opportunities to compose and refine guideline development methodologies, create efficiencies of process, and access committed funds. It was therefore hypothesised that guidelines originating from established programmes would have higher domain scores than those produced outside an established system. To test this hypothesis, a series of one way ANOVAS on quality scores was undertaken for each domain with type of guideline programme (established/not established) as the between subject factor.
 - (b) It can be argued that guidelines supported by well documented technical information—either within the guideline itself or as part of supporting reports or publications—will have domain scores higher than those without this documentation. To test this notion, Kendall's tau B rank correlation tests on quality scores for each domain were undertaken.
 - (c) Guidelines developed as national policies should be particularly robust because of the authority conferred on them. It was therefore predicted that guidelines created on a national level should be of higher quality than regional or local ones. To test this notion a series of one way ANOVAS on quality scores was undertaken for each domain with level status (national/other guidelines) as the between subject factor.
- (3) Criterion validity: as there is no gold standard in this area, participants' overall assessment scores were used as a proxy measure. Assessments of criterion validity were assessed by calculating the Kendall's tau B rank correlation coefficients

Table 1 Domain structure for guideline quality obtained from principal components analysis, mean (SD) values of domain scores, and percentage of variance explained by each domain (item numbers represent the order in the instrument)

	Coefficient*
<i>Domain 1: Scope and purpose</i>	
Mean percentage domain score = 69.3; SD = 21.3; range 16.7–97.2; % variance = 4.6	
1. The overall objective(s) of the guideline is (are) specifically described	0.594
2. The clinical question(s) covered by the guideline is (are) specifically described	0.768
3. The patients to whom the guideline is meant to apply are specifically described	0.702
<i>Domain 2: Stakeholder involvement</i>	
Mean percentage domain score = 36.1; SD = 18.9; range 4.2–68.7; % variance = 6.6	
4. The guideline development group includes individuals from all the relevant professional groups	0.643
5. The patients' views and preferences have been sought	0.580
6. The target users of the guideline are clearly defined	0.683
7. The guideline has been piloted among end users	0.471
<i>Domain 3: Rigour of development</i>	
Mean percentage domain score = 40.7; SD = 25.0; range 0–89.3; % variance = 42.3	
8. The systematic methods were used to search for evidence	0.794
9. The criteria for selecting the evidence are clearly described	0.763
10. The methods used for formulating the recommendations are clearly described	0.750
11. The health benefits, side effects and risks have been considered in formulating the recommendations	0.689
12. There is an explicit link between the recommendations and the supporting evidence	0.753
13. The guideline has been externally reviewed by experts prior to its publication	0.589
14. A procedure for updating the guideline is provided	0.619
<i>Domain 4: Clarity and presentation</i>	
Mean percentage domain score = 65.8; SD = 14.1; range 37.5–91.7; % variance = 8.6	
15. The recommendations are specific and unambiguous	0.716
16. The different options for management of the condition are clearly presented	0.589
17. Key recommendations are easily identifiable	0.739
18. The guideline is supported with tools for application	0.640
<i>Domain 5: Applicability</i>	
Mean percentage domain score = 36.9; SD = 23.2; range 0–91.7; % variance = 6.1	
19. The potential organisational barriers in applying the recommendations have been discussed	0.804
20. The potential cost implications of applying the recommendations have been considered	0.697
21. The guideline presents key review criteria for monitoring and/or audit purposes	0.684
<i>Domain 6: Editorial independence</i>	
Mean percentage domain score 30.3; SD = 22.4; range 0–72.2	
22. The guideline is editorially independent from the funding body	
23. Conflicts of interest of guideline development members have been recorded	New item

*Coefficients from varimax rotated factor matrix.

between the appraisers' domain scores and the overall assessment scores.

RESULTS

The median time for appraising a guideline was 1.5 hours in both field studies. This included reading the guideline and completing the instrument. All appraisals were completed and returned.

Refinement of instrument

Principal components analysis of the draft instrument items yielded a five-factor solution that generally supported the

domains of quality identified in the development phase. Table 1 shows the list of items and their loading (correlation) coefficients on each of the five domains from the rotated factor matrix.

Editorial independence appeared to load on several domains. In response, it was shifted to a sixth domain in the final version of the instrument and a new item addressing conflicts of interest was included. Two items—"The guideline is clearly structured" and "The potential problems with changes of attitude or behaviour of health care professionals in applying the guidelines have been considered"—were removed from the final version of the instrument because of

Table 2 Interrater reliability and internal consistency for each quality domain (n=33)

Domains	Intraclass correlation*				Cronbach α
	1 appraiser	2 appraisers	3 appraisers	4 appraisers	
1. Scope and purpose	0.44	0.61	0.70	0.76	0.88
2. Stakeholder involvement	0.47	0.64	0.72	0.78	0.72
3. Rigour of development	0.71	0.83	0.88	0.91	0.88
4. Clarity and presentation	0.25	0.39	0.49	0.57	0.69
5. Applicability	0.50	0.67	0.75	0.80	0.79
6. Editorial independence	0.34	0.51	0.61	0.67	0.64

*The Spearman-Brown formula to obtain the ICC for the mean of k ratings from the ICC of 1 rating is: $ICC_k = k(ICC_1) / 1 + (k - 1)ICC_1$.

Table 3 Standardised guideline scores and their confidence intervals for each domain according to guideline programme, level of background information, and national policy

	Domain 1	Domain 2	Domain 3	Domain 4	Domain 5	Domain 6
All guidelines (n=33)	69.3 (61.7 to 76.9)	36.1 (29.4 to 42.8)	40.7 (31.9 to 49.6)	65.8 (60.8 to 70.8)	36.9 (28.7 to 45.1)	30.3 (22.3 to 38.2)
Guideline programme						
Developed within a guideline programme (n=20)	68.2 (58.5 to 78.0)	35.6 (27.5 to 43.8)	44.2 (33.1 to 55.3)	66.6 (59.8 to 73.4)	34.9 (24.9 to 44.9)	36.7 (26.5 to 47.0)*
Outside a guideline programme (n=13)	70.9 (57.1 to 84.7)	36.9 (23.8 to 50.0)	35.3 (19.1 to 51.5)	64.4 (56.1 to 72.7)	39.8 (24.0 to 55.7)	20.3 (8.1 to 32.5)
Level of background information						
No information (n=7)	63.5 (42.2 to 84.8)	29.5 (13.8 to 45.1)	23.8 (6.9 to 40.8)	58.6 (43.2 to 74.1)	38.1 (12.2 to 64.0)	30.4 (12.1 to 48.7)
Some information/references (n=10)	67.2 (47.8 to 86.7)	31.1 (15.1 to 47.3)	29.4 (16.5 to 42.4)	64.2 (51.2 to 77.1)	29.4 (17.1 to 41.6)	26.1 (7.0 to 45.3)
Detailed documentation (n=16)	73.1 (64.1 to 82.0)	42.1 (33.4 to 50.8)	55.1 (42.5 to 67.8)**	69.8 (65.3 to 74.3)	41.0 (27.9 to 54.0)	32.8 (21.35 to 44.3)
National policy						
Guidelines developed before (n=13)	71.2 (58.1 to 84.2)	34.2 (22.9 to 45.5)	29.0 (16.3 to 41.6)	67.8 (59.9 to 75.7)	41.8 (25.8 to 57.8)	25.9 (10.6 to 41.1)
Guidelines developed after (n=20)	68.1 (57.9 to 78.2)	37.3 (28.2 to 46.4)	48.3 (36.7 to 60.0)*	64.4 (57.4 to 71.3)	33.6 (23.9 to 43.3)	33.1 (23.6 to 42.7)

*p<0.05, **p<0.01.

failure to establish adequate reliability in the first field test. Finally, 10 items were reworded slightly in the final version of the instrument in response to feedback received from the appraisers (see Face validity below). The refined instrument in its final form contained 23 items grouped into six domains with the 4 point Likert scale to score each item (table 1).

Final instrument properties

Reliability

Internal consistency ranged between 0.64 and 0.88 and was acceptable for most domains (table 2). The lower α coefficient found for domain 6 (editorial independence) was not surprising as this domain was composed of only two items. Table 2 also shows the intraclass correlations for each domain as a function of the number of raters. As would be expected, the number of appraisers evaluating a guideline affected reliability; increasing the number of raters resulted in substantially higher ICCs.

Validity

Face validity

Results from the first field test indicated that the appraisers found the instrument useful to assess guidelines (95%) and the user guide helpful (98%). However, almost half of the participants reported having difficulties with at least one item of the instrument (49%). The most commonly reported problem was that guidelines lacked the detailed information necessary to assign a score. After refinement of the instrument, results from the second field test showed that the percentage of

appraisers reporting difficulties with at least one item in the instrument decreased to 29%.

Construct validity

Tests of the first hypothesis showed that guidelines produced as part of a guideline programme had significantly higher scores on domain 6 (editorial independence) than those published outside a programme ($p<0.05$). Tests of the second hypothesis showed that guidelines with technical documentation had higher scores on domain 3 (rigour of development) than those published without documentation ($p<0.01$). Finally, tests of the third hypothesis revealed that guidelines produced after the publication of a national policy had significantly higher quality scores on domain 3 (rigour of development) than did their counterparts ($p<0.05$). No other significant differences emerged on any of the other domains for any of the contrasts (table 3).

Criterion validity

Kendall's tau B rank correlation coefficients between the appraisers' domain scores and their overall assessments were all highly significant ($p<0.001$), providing some evidence of criterion validity using this proxy measure. Table 4 shows the correlation matrix of the six quality domains. With one exception, the domains tended to be more highly correlated with overall judgement than with each other.

DISCUSSION

This is the first time an appraisal instrument for clinical practice guidelines has been developed and tested at an

Table 4 Correlation between each domain and overall judgement

	Domain 1	Domain 2	Domain 3	Domain 4	Domain 5	Domain 6	Overall
Domain 1	1.00						
Domain 2	0.81	1.00					
Domain 3	0.56	0.71	1.00				
Domain 4	0.56	0.60	0.56	1.00			
Domain 5	0.49	0.55	0.38	0.57	1.00		
Domain 6	0.48	0.56	0.56	0.59	0.49	1.00	
Overall	0.79	0.88	0.87	0.77	0.67	0.74	1.00

Domain 1: Scope and purpose; domain 2: Stakeholder involvement; Domain 3: Rigour of development; Domain 4: Clarity and presentation; Domain 5: Applicability; Domain 6: Editorial independence.

Box 2 Criteria of high quality clinical practice guidelines

1. Scope and purpose

Contain a specific statement about the overall objective(s), clinical questions, and describes the target population.

2. Stakeholder involvement

Provide information about the composition, discipline, and relevant expertise of the guideline development group and involve patients in their development. They also clearly define the target users and have been piloted prior to publication.

3. Rigour of development

Provide detailed information on the search strategy, the inclusion and exclusion criteria for selecting the evidence, and the methods used to formulate the recommendations. The recommendations are explicitly linked to the supporting evidence and there is a discussion of the health benefits, side effects, and risks. They have been externally reviewed before publication and provide detailed information about the procedure for updating the guideline.

4. Clarity and presentation

Contain specific recommendations on appropriate patient care and consider different possible options. The key recommendations are easily found. A summary document and patients' leaflets are provided.

5. Applicability

Discuss the organisational changes and cost implications of applying the recommendations and present review criteria for monitoring the use of the guidelines.

6. Editorial independence

Include an explicit statement that the views or interests of the funding body have not influenced the final recommendations. Members of the guideline group have declared possible conflicts of interest.

international level. Created through a rigorous and iterative process by a collaboration of international experts in clinical guidelines, the instrument was applied to 100 guidelines by over 260 appraisers from 11 countries. Previous studies on similar instruments have been limited to appraisers working in the same institution and from the same country.³⁻⁷ This study resulted in a rigorously developed set of criteria for appraising guidelines (box 2) that can be helpful for clinical practice in two ways: (1) to help clinicians to differentiate between guidelines from different sources, and (2) as a support to the development of high quality guidelines for medical practice.

Our results show that the instrument is sensitive to differences in important aspects of clinical practice guidelines, and it can be used consistently by a wide range of professionals from different cultural backgrounds. Health professionals, policy makers, and consumers were all able to appraise guidelines with the AGREE questions and user guide. The appraisers found the instrument easy to apply and perceived it to be useful for judging the quality of guidelines.

When interpreting the results, several considerations must be kept in mind. Firstly, the factor analysis confirmed our conceptual framework, lending support to the assumption that the quality of clinical guidelines is composed of distinct domains, each assessing key quality attributes. However, the concept of guideline quality is still grounded in assumptions that need testing empirically, and we do not know the relative contribution of each domain to the overall quality of a guideline. Construct validity, based on three a priori hypotheses, was not strong. It was somewhat surprising to observe that

national (as opposed to local) development and established (as opposed to more recent) programmes supporting production did not predict quality more strongly. The high correlations found between the domain scores and the overall assessment corroborated the modest criterion validity, although the effect may be attenuated by the fact that the appraisers made their global ratings after assessing the guidelines.

Secondly, the reliability of the domains is directly affected by the number of appraisers assessing one guideline. Thus, using four appraisers will yield a more reliable assessment than using a single appraiser.¹⁷ In this study average ratings of four raters provided the most reliable assessment and we recommend that at least four raters should be used when using the instrument.

Finally, we were not able to demonstrate conclusively the validity of our instrument. The instrument assesses the methodological quality of a guideline and this relies heavily on how well documented the guideline development process is.¹⁸ However, explicit reporting does not guarantee optimal recommendations. A well reported guideline may contain flawed recommendations and, conversely, an unsystematically constructed one may provide sound evidence.¹⁹ Nevertheless, the criteria we used are accepted as key determinants of valid and effective guidelines among methodologists, and the domains are quite clear. Validation of the instrument is a challenging task. We are currently undertaking detailed content analysis of the appraised guidelines as part of our research programme. This will provide a separate measure of construct validity.

AGREE has considerable implications for research and policy. These standards for the development and reporting of clinical practice guidelines can be used by guideline producers worldwide. The adoption of such standards can improve the consistency and quality of the reporting process.²⁰ The sharing of standards across countries will facilitate international comparison of guidelines and can provide a framework for studies aimed at understanding why guidelines for the same condition may produce differing recommendations.²¹⁻²²

As the number of clinical practice guidelines submitted for publication increases, there is a need to ensure that they satisfy certain minimum requirements. AGREE can be adopted by editors of peer reviewed journals as a framework to assess the quality of clinical guidelines in the same way that CONSORT is used to judge the quality of randomised controlled trials and meta-analyses.²³⁻²⁴

Given the expansion of national guideline programmes, governments and other agencies must ensure the guidelines are of the highest quality before they endorse them or promote their use in practice. Furthermore, as international cooperation between countries grows there is a strong incentive for policy makers to develop a concerted approach to quality management initiatives, including clinical practice guidelines. The AGREE instrument can enhance this process. This is already taking place as several agencies—such as the National Institute for Clinical Excellence (NICE) in the UK, the National Federation of Cancer Centres (FNCLCC) in France, The Agency for Quality in Medicine in Germany (ÄZQ), and the Scottish Intercollegiate Guidelines Network (SIGN)—are using AGREE in the context of their guidelines programme. The World Health Organisation has adopted the AGREE instrument to assess its guidelines.

In conclusion, the AGREE collaboration has developed an instrument for guideline appraisal using a rigorous methodology. The instrument has been applied to different clinical practice guidelines in 11 countries by a large number of appraisers from a variety of backgrounds. We recommend that guideline producers use this instrument while planning their programmes, and potential guideline users use it to evaluate the quality of guidelines before adopting them.

The AGREE instrument is available on the AGREE website (www.agreecollaboration.org).

Key messages

The problem

Clinical practice guidelines are used increasingly by government agencies and professional organisations around the world to improve patient care, but many published guidelines do not meet the basic quality criteria. There is a pressing need for internationally recognised criteria to assess guidelines that are valid and reliable.

What this study adds

- An international collaboration, the AGREE Collaboration, has developed an instrument for assessing the process of guideline development that is reliable and is acceptable in European and non-European countries.
- It was not possible to confirm the validity of the instrument.
- The instrument provides common standards to improve the quality process and reporting of guideline development worldwide.
- These standards can be used for the planning, execution, and monitoring of guideline programmes and for comparing guidelines internationally.

ACKNOWLEDGEMENTS

The authors would like to thank the 264 appraisers from the 11 countries who participated in the study and the following colleagues for their valuable comments on the first draft of the instrument: Richard Baker, Martin Eccles, Roeland Geijer, Trisha Greenhalgh, Allen Hutchinson, Nick Hicks, Chris Silagy, Siep Thomas, Richard Thomson, Michel Wensing and Steven Woolf.

Writing group: Françoise Cluzeau (FC), St George's Hospital Medical School, London, UK; Jako Burgers (JB), University of Nijmegen, The Netherlands; Melissa Brouwers (MB), McMaster University and Cancer Care Ontario, Hamilton, Ontario, Canada; Richard Grol (RG), University of Nijmegen/University of Maastricht, The Netherlands; Marjukka Mäkelä (MM), Finnish Office for Health Care Technology Assessment, Finland; Peter Littlejohns (PL), National Institute for Clinical Excellence, London, UK; Jeremy Grimshaw (JG), Health Services Research Unit, University of Aberdeen, UK; Claire Hunt (CH), Institute of Psychiatry, London, UK.

FC, JB, RG and PL developed the first draft of the instrument and designed the field study. FC and JB drafted the paper and undertook the analyses with CH. MB, MM, RG, PL and JG helped write the final draft.

Contributors: The following individuals provided input into the design and field testing of the AGREE instrument and commented on earlier drafts of the paper: José Asua, Basque Office for Health Technology Assessment, Spain; Anne Bataillard, Fédération Nationale des Centres de Lutte Contre le Cancer, Paris, France; George Browman, Hamilton Regional Cancer Centre, Hamilton, Canada; Bernard Burnand, Institut Universitaire de Médecine Sociale et Préventive, Lausanne, Switzerland; Pierre Durieux, Hôpital Européen Georges Pompidou, Paris, France; Béatrice Fervers, Fédération Nationale des Centres de Lutte Contre le Cancer, Paris, France; Roberto Grilli, Agenzia Sanitaria Regionale, Bologna, Italy; Steven Hanna, McMaster University, Hamilton, Ontario, Canada; Pieter ten Have, Utrecht, The Netherlands; Albert Jovell, Fundacio Biblioteca Josep Laporte, Barcelona, Spain; Niek Klazinga, Academisch Medisch Centrum University of Amsterdam, The Netherlands; Finn Kristensen, Danish Institute for Health Technology Assessment, Copenhagen, Denmark; Pia Bruun Madsen, Danish Institute for Health Technology Assessment (DITHA), Copenhagen, Denmark; Juliet Miller, SIGN (Scottish Intercollegiate Guidelines Network), Edinburgh, UK; Günter Ollenschläger, Agency for Quality in Medicine, Cologne, Germany; Safia Qureshi, SIGN (Scottish Intercollegiate Guidelines Network), Edinburgh, UK; Rosa Rico-Isturiz, Basque Office for Health Technology Assessment, Spain; John-Paul Vader, Institut Universitaire de Médecine

Sociale et Préventive, Lausanne, Switzerland; Joost Zaat, Centre for Quality of Care Research, The Netherlands.

Funding: The research was funded by a grant from the EU BIOMED2 Programme (BMH4-98-3669). The work in Switzerland was funded from the Swiss Federal Office for Education and Science (OFES 97.0447). The Health Services Research Unit, University of Aberdeen is funded by the Chief Scientist Office of the Scottish Executive Department of Health. The views expressed are those of the authors and not the funders.

Conflict of interest: none.

REFERENCES

- 1 Woolf SH, Grol R, Hutchinson A, *et al.* Potential benefits, limitations, and harms of clinical guidelines. *BMJ* 1999;**318**:527–30.
- 2 Woolf SH, Grol R, Hutchinson A, *et al.* An international overview. In Eccles MP, Grimshaw JM, eds. *Clinical practice guidelines*. Oxford: Radcliffe Medical Press, 2000.
- 3 Field MJ, Lohr KN, eds. *Guidelines for clinical practice. From development to use*. Institute of Medicine. Washington DC: National Academy Press, 1992.
- 4 Shekelle PG, Woolf SH, Eccles M, *et al.* Clinical guidelines: developing guidelines. *BMJ* 1999;**318**:593–6.
- 5 Grimshaw JM, Russell IT. Achieving health gain through clinical guidelines. I: Developing scientifically valid guidelines. *Qual Health Care* 1993;**2**:243–8.
- 6 Grilli R, Magrini N, Penna A, *et al.* Practice guidelines developed by specialty societies. The need for a critical appraisal. *Lancet* 2000;**355**:103–6.
- 7 Shaneyfelt TM, Mayo-Smith MF, Rothwangl J. Are guidelines following guidelines? The methodological quality of clinical practice guidelines in the peer-reviewed medical literature. *JAMA* 1999;**281**:1900–5.
- 8 Lohr KN, Field MJ. A provisional instrument for assessing clinical practice guidelines. In: Field MJ, Lohr KN, eds. *Guidelines for clinical practice. From development to use*. Washington DC: National Academy Press, 1992.
- 9 Cluzeau F, Littlejohns P, Grimshaw J, *et al.* Development and application of a generic methodology to assess the quality of clinical guidelines. *Int J Qual Health Care* 1999;**11**:23–8.
- 10 Grol R, Dalhuijsen J, Thomas S, *et al.* Attributes of clinical guidelines that influence use of guidelines in general practice: observational study. *BMJ* 1998;**317**:858–61.
- 11 Thomson R, Lavender M, Madhok R. How to ensure that guidelines are effective. *BMJ* 1995;**311**:237–42.
- 12 Lohr KN. The quality of practice guidelines and the quality of health care. In: *Guidelines in health care*. Report of a WHO Conference, January 1997, Baden-Baden: Nomos Verlagsgesellschaft, 1998.
- 13 MacCallum RC, Widaman KF, Zhang S, *et al.* Sample size in factor analysis. *Psychol Methods* 1999;**4**:84–99.
- 14 Norman GR, Streiner DL. *Biostatistics. The bare essentials*. 2nd ed. St Louis: Mosby, 2000.
- 15 Bland JM, Altman DG. Cronbach's alpha. Statistics notes. *BMJ* 1997;**314**:572.
- 16 Fleiss J.L. The measurement of interrater agreement. In: *Statistical methods for rates and proportions*. New York: John Wiley & Sons, 1981.
- 17 Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;**86**:420–8.
- 18 Hayward RS, Wilson MC, Tunis SR, *et al.* Users' guides to the medical literature. VIII. How to use clinical practice guidelines. A. Are the recommendations valid? The Evidence-Based Medicine Working Group. *JAMA* 1995;**274**:570–4.
- 19 Moher D, Ba P, Jones A, *et al.* Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet* 1998;**352**:609–13.
- 20 Cluzeau F, Littlejohns P. Appraising clinical guidelines in England and Wales. The development of a methodological framework and its application to policy. *Jt Comm J Qual Improve* 1999;**25**:514–21.
- 21 The AGREE Collaborative Group. Guideline development in Europe: an international comparison. *Int J Technol Assess Health Care* 2000;**16**:1039–49.
- 22 Grol R, Eccles M, Maisonneuve H, *et al.* Developing clinical practice guidelines: the European experience. *Disease Management Health Outcomes* 1998;**4**:255–66.
- 23 Moher D, Schulz KF, Altman D for the CONSORT Group. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *JAMA* 2001;**285**:1987–91.
- 24 Moher D, Jones A, Lepage L for the CONSORT Group. Use of the CONSORT statement and quality of reports of randomized trials. A comparative before-and-after evaluation. *JAMA* 2001;**285**:1992–5.